TEST Engineering & Management

Jan-Feb 2020
ISSN: 0193-4120 Page No. 17609 - 17617

An approach to analyze and predict highway accident scenarios in India

SD Chitnis[1], P Gokhale[2], SS Sikhakolli[3]

1.Research Scholar, Science and Technology Department, Vishwakarma University, S.N. 3/6, Laxminagar, Kondhwa- Pune.

2. Associate Professor, Science and Technology Department, Vishwakarma University, S.N. 3/6, Laxminagar, Kondhwa- Pune.

3. Assistant Professor, Science and Technology Department, Vishwakarma University, S.N. 3/6, Laxminagar, Kondhwa- Pune.

Abstract: Road traffic is increasing at a huge rate which leads to a large number of accidents in India. The average number of vehicles in India is growing at the rate of 10.16% annually, over the last few years [1]. Furthermore, the frequencies of road traffic accidents (RTA) also differ from time to time in the same location causing heterogeneous data. However, this research paper addresses the statistical analysis and machine learning approach to work on this type of data and shows how it plays a major role in recognizing the causes. Data mining techniques such as regression is widely used in the analysis of road accident data. The innovative approach like machine learning-based predictive analytics like regression to foresee the number of accidents that may happen shortly. It can extract knowledge from complex data without relying on a prior underlying relationship between data variables. The paper is divided into three major sections: In the first section, a theoretical concept with an overview of literature survey and dataset is elaborated, the second part presents the methodologies to be implemented and finally, the third section describes the analysis using a statistical model and a regression model with results and discussion.

Keywords: RTA, heterogeneous data, statistical analysis, Data Mining, Machine Learning, regression, black spot

Introduction

Road traffic accidents (RTA) are undistinguishable incidents in India. Chennai leads the pack with a total of 7486 accidents followed by Delhi and Bengaluru [2]. The maximum deaths are recorded in Delhi followed by Chennai. According to the World Health Organization, there were 1.25 million road traffic deaths globally in 2013. Alcohol and other drugs are found to be the most contributing cause in up to 22% of vehicular accidents on the world's highways and byways [3]. RTA accidents lead to numerous hazards like lifetime disability or death, and monetary cost to an individual as well as to the society. According to official statistics, 0.11 million deaths occurred in India due to road traffic accidents in 2006, which is nearly 10% of the total road traffic deaths in the world [3]. 1214 road crashes occur every day in India. Two-wheeler accounts for 25% of total road crash deaths. 20 children under the age of 14 die every day due to road crashes in the country. 377 people die every day, equivalent to a jumbo jet crashing every day [4].

In India, in-discipline especially among the young drives is more susceptible to RTA. Numerous factors that increase the risk of collision includes road features and road condition, nature, and causes of an accident, whether the condition and behavior of a driver. In this paper, emphases are on these attributes as an accident will not only risk a person's own life but may also root cause an incident life to be lost.

The problem of RTA must be approached by addressing 4 W's: Where Who, What, and Why? Where and with Whom is it concerned? – Commuters in urban as well as rural areas are

Jan-Feb 2020
ISSN: 0193-4120 Page No. 17609 - 17617

prone to accidents. What would happen if the said problem is not sorted? – According to the Ministry of Road Transport Highway, in 2004, it was the 9th leading cause of death but if this situation is not controlled, then by 2030 it will be the 5th leading cause of death [5]. Why is it crucial that this is solved? – India has around 600 million young people and they are set to change the world [7], but at the same time, India has the highest death-rate of youth because of road accidents. This is the reason why we need to control the situation so as not to affect our nation at a great level.

Generally, the segmentation of RTA is based on expert domain knowledge. Although expert knowledge can lead to a segmentation of data, it does not assurance that each segment consists of a homogenous group of traffic accidents. Therefore, analysis of RTA could assist from a data analysis technique that helps in the course of traffic accident segmentation. Such techniques can be found in the area of data mining. Data mining uses many different techniques and algorithms to discover the relationship in a large amount of data. It is the process of employing one or more computer learning techniques to analyze automatically. From the database, these techniques from contained data can extract knowledge. Road accidents happen on highways such as on signals, in crowded areas due to the wrong side and speedy driving, and inappropriate infrastructure of local roads. The analysis of such type of heterogeneous data is possible using Machine learning algorithms.

Machine Learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed [6]. This can be done by acquiring knowledge from previous data.

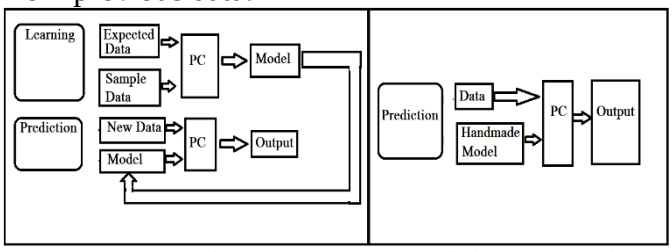


Figure (a): Machine Learning Model - Figure (b): Traditional Model

Figures (a) and (b) show that in what way does a machine learning is different from the traditional one. In this, the learning phase output model is one of the inputs in the prediction phase along with 'new data' to get the output, unlike traditional modeling. It divides our task into 2 phases i.e. learning and prediction.

Machine Learning is divided into two micro-areas: 1] Supervised Learning: The right answer should be given, but the machine needs to find the more correct answer. In this type of learning, the labeled data would be given. Example: Regression and classification. 2] Unsupervised Learning: The right answer would not be given and the data is unlabeled. Example: Clustering and dimension reduction.

Most studies used statistical techniques [8, 9], data mining techniques [10] along with machine learning algorithms to analyze the road accident data. The accident rates were found to be significantly related to road design parameters of study as well as the nature of accidents. Many research studies incorporating different aspect of RTAs have been carried out by different workers at different times across the globe [13, 14, 15, 16, 17, 18]. However, realizing the need to establish baseline information on RTAs, the present study was conducted in this part of India with the following objectives:

Objectives

- To analyze road traffic accident cases using a predictive model to identify Black-spot, and
- To find whether the only the causes of the accident can define the accident classification type i.e. major or minor and to identify the probable sources of crisis i.e. minor or major

Literature Review

The proposed research review takes an overview of various statistical methods and machine learning algorithms like regression analysis, k-means clustering algorithms, support vector machine (SVM), k-modes, Latent Class Clustering (LCC) which are used in the analysis of road accident data. In this research work, various methods used by different researchers are

Jan-Feb 2020
ISSN: 0193-4120 Page No. 17609 - 17617

reviewed to propose a better technique, to enhance road safety and to curb the RTA ratio.

The paper entitles "A Statistical Analysis of Road Traffic Accidents in Dibrugarh City, Assam, India", by Ajit Goswami and et.al (2011) collected the data from police records and case diaries of Dibrugarh Police station were studied [4]. Ky-plot and SPSS software with bivariate comparisons were used. The data is analyzed statistically by data interpretation using the Degree of freedom, Chi-square test, and Kruskal-Wallis test.

The results obtained in this paper states that human characteristics such as rush and pure negligence make 95.38% of the total RTAs. During the day time, i.e. from 6 am to 6 pm, 60% of the accidents were recorded and the peak time was between 12 noon to 6 pm, when 38.4% of the accidents were recorded. The author of this paper concludes that road traffic accidents are avoidable and only a more integrated approach is expected across many sectors and many disciplines. He admits that fewer data on accident reports at police stations are revealing of lack of awareness of road traffic accident reporting. As mentioned, fewer data may not able to give more precise results.

The paper entitled, "A data mining framework to analyze road accident data", by Sachin Kumar [5] and et.al in the year 2015 focused on accident data analysis associated with road traffic accidents. The data set consists of more than 11,000 road accident records from 2009 to 2015 i.e. 6 years, in Dehradun District of Uttarakhand State. As per the researcher, road accident data makes the analysis task difficult as it is heterogeneous. Data segmentation has been used extensively to overcome this sort of problem of heterogeneity of the accident records. In this paper, Clustering analysis is a proposed framework that uses the K-modes clustering technique in combination with Latent Class Clustering (LCC) followed by association rule mining. This data can be grouped into different homogenous segments. It helps in removing heterogeneity to some extent in the road accident data. Thus, association rule mining is further applied to a cluster as well as on the entire data set (EDS) to generate more appropriate rules. They included and worked on 11 attributes. Using Apriori algorithm accident-prone circumstances can be identified and trend analysis can be performed for each cluster and on EDS. We are working on the three main attributes which are forbidden by the researcher those are: vehicle condition, driver profile, and road conditions.

The paper entitled "Road Traffic Accidents in India: Issues and Challenges" by Sanjay Kumar Singh [6] was published in the year 2017. This research paper focused on analyzing road accidents at the national, state, and metropolitan city levels of India. According to the analysis, the number of injuries and fatalities in road accidents varies according to age, gender, month, weather, and time. The age group 30-59 years shows higher levels of fatality and wounds in men rather than women. Although the burden of road accidents in India is slightly lower in their metropolitan cities, about 50% of the cities show a riskier relative to the countryside.

The paper states that out of the total population of India, 20% of people comprise in the age group of 30-44. However, 35% of total road accident fatality occurs in this group. As mentioned in the paper most of the road accidents occur at the wee hours, therefore we need to find out the root cause and give the appropriate solutions. Driver's fault is one of the major reasons for 78% of total accidents.

Database

Nature of data:

For this, apt and more precise information is required for the data analysis to get precise results.

The present study is based on secondary sources of data i.e. data collected from the National Highways Authority of India, NHAI-Pune, India. The official records were available from 2011 – 2017 [19]. The dataset contains 2368 records and 14 attributes. The data description can be found in the inception report document of the NHAI Analytic Reference Guide, 2011 to 2017 [20]. A statistical technique like a Chi-Square test and Regression Analysis was applied. A complete enumeration of data was done. Besides, other relevant information was collected from the

Jan-Feb 2020
ISSN: 0193-4120 Page No. 17609 - 17617

concerned officials through interviews and personal discussions. The dataset consists of 2504 road accidents for 7 years period from 2011 to 2017 of Pune-Satara NH-4 highway locations. After pre-processing, 2368 accidents records have been considered for this research.

This research study emphases on the following attributes: Accident Location (Location), Causes of Accident (CoA), Nature of Accident (NoA), Classification of Accident-Accident Type (CLoA), Road Features (RF), Road condition (RC), Intersection Type (INTT), and Whether Condition (WC).

Model Flow

Following Figure(c) shows the flow of Machine Learning Model-

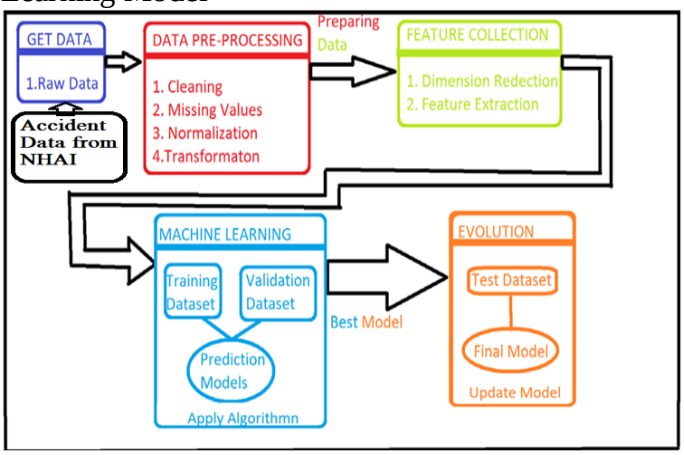


Figure (c): Machine Learning Model Flow

As shown in Figure (c), get the RAW data followed by Data Preprocessing. In the machine learning model, learning includes pre-processing, in which we clean the raw data collected from NHAI. Pre-processing minimizes unwanted data for a better-processed result. The collected data can be preprocessed and filtered. The standardization of data is done by using the formula i.e.,

$$X_{standardization} = \frac{X - mean(X)}{Standard\_deviation(X)}$$

Data pre-processing as well as feature extraction has a significant impact on the model performance. If the number of features becomes similar or bigger than the number of observations stored in a dataset, then this can most likely lead to Overfitting. To avoid this type of problem, it is necessary to apply dimensionality reduction techniques. In feature collection, feature extraction can reduce the number of features in a dataset by creating new features from the existing ones. After this, in our model machine learning algorithm like regression plays a major role. In this, the training data is used to make sure the machine recognizes patterns in the data. The cross-validation data is used to ensure better accuracy and efficiency of the regression algorithm, which is used to train the Machine. Whereas, test data is used to see how well the machine can predict new answers based on its training. This testing of data is done to get the best final model.

Methodologies

i. Statistical Analysis: Data were analyzed using the Chi-square test and P-value below 0.05 was considered as statistically significant.

The Chi-square test is also known as a goodness of fit in statistics. It is intended to test and measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent. It is commonly used for testing the relationships between the variables or attributes.

Data Interpretation: The following methods were applied to analyze the data - [12]

a) *Degree of freedom (d.f.):* The number of independent variants that make up the statistic (e.g. χ^2) is known as the degree of freedom (d.f.). The number of degrees of freedom, in general, is the total number of observational less the number of independent constraints imposed on the observations. For example, if n1 is the number of independent constraints in a set of data of n observations then d.f. = (n-n1). Therefore, in a set of 'n' observations usually, the degree of freedom for χ^2 is (n-1), one d.f. being lost because of the linear constraint $\sum_i Oi = \sum_i Ei = N$, on the frequencies. If 'r' independent linear constraint one imposed on the cell frequencies, then the d.f. reduced by 'r'. Besides, if any of the population parameter(s) is calculated from the given data and used for computing the expected frequencies then in applying χ^2-test of the goodness of fit, we have

TEST Engineering & Management

Jan-Feb 2020
ISSN: 0193-4120 Page No. 17609 - 17617

to subtract one d.f. for each parameter calculated. Thus if's' is the number of population parameters estimated from the sample observation (n in number), then the required number of degree of freedom for χ^2 -test is (n-s-1). If anyone or more of the theoretical frequencies are less than 5 then in applying χ^2-test we have also to subtract the degrees of freedom lost in pooling these frequencies with the preceding or succeeding frequency. In a(r × s) contingency table, in calculating the expected frequencies, the row totals, the column totals remain fixed. The fixation of 'r' column totals and 's' row totals imposes (r+s) constraints on the cell frequencies [12]. But since $\sum_{i=1}^{r}(Ai) = \sum_{j=1}^{s}(Bj) = N$, the total number of independent constraints is only (r+s-1) (Table 1). Further, since the total number of the cell frequencies is r+s, the required number of degrees of freedom is $v = rs - (r + s - 1) = (r - 1)(s - 1)$ [12]

b) *Chi-square test for goodness of Fit:* Formulated by Prof. Karl Pearson in 1900, it is a very powerful test for testing the significance of the discrepancy between theory and experiment. It enables us to find if the deviation of the experiment from theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data. If O_i, (i= 1, 2,. . . n) is a set of observed (experimental) frequencies and E_i (i=1, 2, . .,n) is the corresponding set of expected (theoretical or hypothetical) frequencies, then Karl Pearson's Chi-square statistic given by, [12]

$$\chi 2 = \sum_{i=1}^{n}\left[\frac{(O_i - E_i)^2}{E_i}\right] \ldots (1.1) \ (\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i)$$

-follows Chi-square distribution with (n-1) d.f.

c) *χ^2 – test for independence of attributes:* [12] The χ^2 - test of independence of attributes is also applied to test the independence of season and accident, and also to test the independence of hours of the day and the accident. The test is as follows -

Let us consider two attributes A and B, A divided into r classes A_1, A_2,... ., A_r and B divided into s classes B_1, B_2,…, B_s. Such a classification in which attributes are divided into more than two classes is known as manifold classification. The various cell frequencies can be expressed in the following table known as (r×s) manifold contingency table where (Ai) is the number of people possessing the attributes A_i, (i=1, 2,…, r), (B_j) is the number of people possessing the attribute B_j (j=1,2,…,s) and (A_i B_j) is the number of people possessing both the attributes A_i and B_j [i= 1, 2, …, r; j = 1, 2, …, s] [12]. Also, $\sum_{i=1}^{r}(A_i) = \sum_{j=1}^{s}(B_i) = N$, is the total frequency. Finally, the exact test for the independence of attributes is very complicated but a fair degree of approximation is given, for large samples (large N), by the chi-square test of goodness of fit, viz. $\chi 2 = \sum_{i=1}^{n}\sum_{j=1}^{s}\left[\frac{\{(AiBj) - (AiBj) \quad \theta\}^2}{(AiBj)\theta}\right]$, which is distributed as a χ^2 - variate with (r-1) (s-1) d.f. Reject the null hypothesis if calculated χ^2 > tabulated χ^2 at 5 % level of significance [16]

In this research, testing of the relationship between the 'Nature of Accident' and the 'Classification of Accident'. In this paper, to find out whether the nature of accident i.e. overturning, rear-end collision, head-on collision, etc. can define the Classification or 'accident type', i.e. the accident is critical (major) or non-critical (minor). We have to find out here that can only one attribute define this accident type or more than one can interpret this. In other words, anyone attribute can suffice to interpret accident type or not. Subsequently, the null hypothesis is that no relationship exists on the categorical variables in the population; they are independent. Following this, the research hypotheses are:

H_0 = The nature of accidents defines accident type or classification. Whereas, the alternate hypothesis is H_1 = The nature of accidents does not define accident type or classification.

The following methods were applied to analyze the data - χ^2 – test for independence of attributes: The χ^2- test of independence of attributes is also applied to test the independence of NoA and CLoA, and also to test when adding one more attribute i.e. RF with NoA and CLoA (accident type).

The test is as follows - Let us explore two attributes Nature of Accident (NoA) and Classification of Accident (CLoA). NoA divided into classes 1. Drunken 2. Overspeeding 3. Vehicle out of Control 4. The fault of a driver of

Jan-Feb 2020
ISSN: 0193-4120 Page No. 17609 - 17617

motor vehicle 5. The fault of a driver of another vehicle, 6. The fault of Pedestrian 7. The fault of Passenger 8. Defect in the mechanical condition of motor vehicle 9. Road Condition 10. Other (Specify) and CLoA divided into 1. Fatal 2. Grievous injury 3. Minor Injured 4. Non-injury.

Cross-Tab: Python- Table of Observed Values						Total
Nature of Accident		Classification of Accident				
		1-Fatal	2-Grievous	3-Minor	4-Non-injury	
Overturning	1	48	171	124	86	429
Head-on	2	78	158	122	57	415
Rear-end	3	99	284	180	98	661
Sideswipe	4	1	6	5	9	21
Right-angled	5	2	27	10	10	49
Skidding	6	48	266	184	91	589
Right-turned	7	1	14	12	10	37
Others	8	46	62	32	27	167
Total:		323	988	669	388	2368

Table (a): Observed Values

Table of Expected Values		Classification of Accident			
Nature of Accident		1 Fatal	2 Grievous	3 Minor	4 Non-injury
Over turning	1	58.5	179.0	121.2	70.3
Head On	2	56.6	173.2	117.2	68.0
Rear End	3	90.2	275.8	186.7	108.3
Side Swipe	4	2.9	8.8	5.9	3.4
Right Angled	5	6.7	20.4	13.8	8.0
Skidding	6	80.3	245.8	166.4	96.5
Right Turned	7	5.1	15.4	10.5	6.1
Others	8	22.8	69.7	47.2	27.4

Table (b): Expected values

Degree of freedom = 21
Level of significance, alpha = 0.05
Tabular value for 21, χ^2 for 0.05 = 32.671
And χ^2 calculated = 90.12.
Thus, χ^2 calculated > χ^2 tabular (Yes), Therefore, the Null hypothesis is rejected.
Then the alternate hypothesis is, the nature of the accident does not define the classification type, e.g. the accident is caused due to drunken, over speeding, etc. *nature of accident* does not classify the accident-type i.e. whether it is a major accident or minor. It indicates that the probable sources of crisis (minor or major) do not depend ONLY on the nature of the accident. In other words, more parameters or attributes are required to define the type of accident.
Thus, when we add one more feature (attribute) i.e. road feature, then the table obtained after cross-tab can be summarized as follows:

Zone	Nature of Accidents + Road Features	Percentage
RED	Overturning + 2,3-Lanes	56%
RED	Sideswipe + 2,3,4-Lanes	24.6%
ORANGE	Head-on, Rear-end Collision + All lanes AND Sideswipe + 1-Lane	15.7%
GREEN	Right-angled, Right-turn Collision, Skidding + 2,3-Lanes	1.6%

Table (c): NoA+RF contributes towards Road Accident

Discussion and Conclusion

Analysis of qualitative data gathered during the present study summarizes the principle factor viz. human. It is a significant contributor to the occurrence of RTAs on NH-4. In the case of road traffic accident data, an association rule mining can identify the several values of the attribute, which are responsible for accident occurrence.
When we add ONE-more attributes, we come to certain Conclusion like -
Table (c) showing three zones, viz. RED, ORANGE, and GREEN by following the intensity of the accidents. The present study recorded 56% of the accidents on 2,3-Lanes while Overturning. Whereas, sideswipe along with overturning which comes under Nature of Accident (NoA), on 2-lanes as well as 3-lanes of Road Features, causes more than 80% of the accidents. However, Right-angled, Right-turn Collision, Skidding on 2,3-lanes, the percentage of an accident is negligible i.e. 1.6%. This concludes the possible and probable sources of CRISIS, which is one of our Objectives. When we apply this on date field and causes of accident i.e. over-speed or drink and drive field, we can set MORE rules and can predict more precisely. It is

Jan-Feb 2020
ISSN: 0193-4120 Page No. 17609 - 17617

also likely that while the number of fatalities reduces in a scenario of overturning, the number of grievous and minor injuries may still rise, i.e. 11% more than the fatal accident cases.

ii. Regression Analysis:

Regression models are coming under the Supervised learning approach. Regression analysis estimates the relationship between the dependent data and independent data. Independent variables are used to predict the value of the dependent variables. It is also known as regressor variables or predictors. When you have more than ONE independent variable like in our accident case and ONE dependent variable, then it is suggestable to use Multiple linear or Multivariate regression.

The RTA prediction model was developed by using 'Multiple Linear Regression Analysis'. Regression models which are a 'supervised learning' approach are used to estimate the relationship between the dependent and independent data. Its most common use is in forecasting or prediction. It is used to 'find out', which factor has the highest impact on the predicted output. It is also used to find out how different variables relate to each other.

Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable [11]. A regression analysis of the number of total accidents divided by accident location (dependent variables) with the selected independent variables was performed. In this work, the scope is to present the accident scenario in Pune-Satara Highway (NH-4) by road traffic accident model based on regression analysis. The importance of normal distributions is undeniable when applying regression models, because interpretation and inferences may not be reliable or valid when the normality assumption is violated.

When you choose to analyze your data using multiple regression, some part of your process involves checking to make sure that the data can be analyzed using multiple regression.

Statistics	NoA	CLoA	CoA	RF	RC	Location
Count	2368	2368	2368	2368	2368	2368
Mean	3.87	2.88	2.51	2.53	2.03	785.32
Std	2.53	1.82	1.14	0.98	1.4	37.54
Min	1	1	0	0	0	725
Max	10	10	10	5	8	870.7

Table (d): Statistics of the attributes

For this, assume that your dependent variable should be measured on a continuous scale i.e. ratio or interval variable. This assumption includes the date (measured in days), road features and conditions (measured for area and type of locations), classification of accidents (measured in types as major or minor), and so forth. The second assumption is if you've 2 or more independent variables, which can be either continuous (NoA, RF, or RC) or categorical (or nominal like CLoA or WC). Another assumption is there needing to be a linear relationship between (i) the dependent variable and each of your independent variables like in our case, and (ii) the dependent variable and the independent variables collectively. There are various ways to check for this linear relationship like creating partial regression plots using the matplotlib library of Python as shown. Finally, errors or residuals are approximately normally distributed need to be checked. The methods to check this assumption are (i) a histogram with a superimposed normal curve and a normal P-P plot or (ii) a normal Q-Q plot as shown in the following Figure (d).

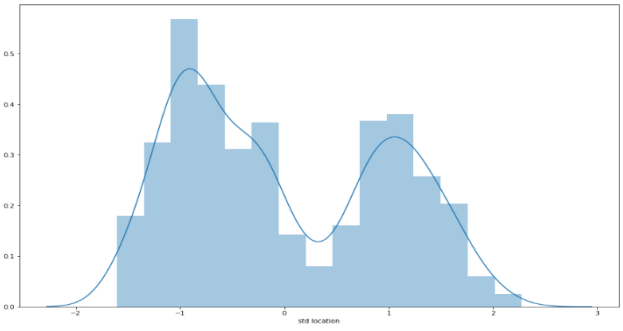
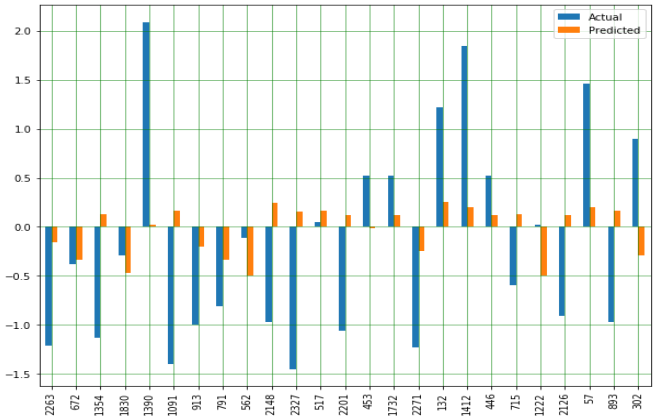


Figure (d): Multiple regression

The above Figure shows some negative skewness along with a fat tail on the right of the first curve.

Jan-Feb 2020
ISSN: 0193-4120 Page No. 17609 - 17617



Graph (1): Difference between Actual and Predicted Values

In multivariable linear regression, the regression model has to find the most optimal coefficients for all the attributes. Following coefficients our regression model has chosen:

Attributes	Coefficients
NoA	-0.090126
CLoA	0.089371
CoA	-0.040058
RF	0.069828

Table (e): Coefficients of a regression model

Above table shows the Nature of accident (rear-end, head-on, etc.), Causes of accidents (drunken, Overspeed, etc.) are negative values. These negative values interpret that to identify the black spot, the attributes like drunken or Overspeed does not always give the correct prediction. As we know that due to drunken case or Overspeed, s/he met with an accident at any location and not at a pre-defined location known as BLACK CORRIDOR. However, Classification of Accident (fatal, minor injury, etc.) and road features (1-lane, 2-lane, etc.) are concerned, coefficients show the positive values. This concludes that these attributes are the more effective and play a major role to identify the BLACK SPOT.

The objectives of regression analysis were to identify: (1) the probable sources of the crisis behind the black spots; (2) the attributes of each accident inside the black spot zones (like in above example NoA and RF); (3) the attributes of the region delimited by the black spots (area, population, demographic density and road network density); and (4) the number and spatial locations of the entities considered in this study as the intersection type (especially joining the village and highway) inside the black spot zones.

Discussion and Conclusion

Multivariate regression analysis comes with the following values:

Mean Absolute Error (MAE) : 0.86

Mean Squared Error (MSE) : 0.98

Root Mean Squared Error (RMSE): 0.99

As we can observe here that our model has returned reasonably good prediction results. The performance of the algorithm can be evaluated by observing the values of the MAE, MSE, and RMSE. Here, RMSE is the standard deviation of the residuals i.e. prediction errors. It tells us how concentrated the data is. You can see that the value of RMSE is 0.99, which is slightly greater than 10% of the value of 'mean' which is 1.86 (i.e. 0.19). This interprets that whatever assumptions we made that this data has a linear relationship is correct. And the features we used have had a high enough correlation to the values we were trying to predict. This means that our algorithm can make reasonably good predictions about the accidents. Only the thing is that we need to have more data to get the best possible prediction. Due to this factor only, RMSE is slightly greater than the mean. On the whole, our predictions are correct.

To summarise the results of the regression model, the classification of accidents (accident type) is positively associated with attributes like the nature of accident and road features and negatively associated with one attribute like nature of the accident. The regression model, on the other hand, estimates an independent effect of an attribute.

The effect of national highways is most counterintuitive and is expected to be in the opposite direction. Since the length of other road types has not been included, NH may be indicating the effect of the overall road network. An increase in the density of the road network may be an indicator of higher congestion. The effect of the urban population indicates higher safety resulting from slower travel speed within

Jan-Feb 2020
ISSN: 0193-4120 Page No. 17609 - 17617

urban areas, as opposed to the faster-moving traffic on rural inter-city roads.

This study has strengths as a societal regression model as well as a statistical model was developed to understand the relationship between accident classification-type and nature of accident with road features in India. This paper reports the use of NHAI data to develop an injury prediction model accounting for exposure of NH (National Highway) users. In the Indian scenario, where the mode of transportation is complex, this study adds a significant understanding of how road death burden/ injury will evolve as travel patterns change in the future. This type of statistical and machine learning (ML) study can be applied on NH-4 (Pune-Satara Highway) in Maharashtra, India, and also can be applied to model injuries at the city or district level.

Reference

[1] A. Lamsal, S.Anand, M. Walia, A. Choudhury, A. Anand (2013), 'Automotive traffic information systems for India'

[2] Usha V.Sagar, 'Cities with maximum Road Accidents', https://www.mapsofindia.com,

[3] Mandip Kumar Nar, Prabhjeet Singh, Paprinder Singh (2019), 'Concept of Internal Momentum Absorber Structure to Reduce Impact During Accidents', *Springer Nature*, pp.144-149

[4] Global Status Report on road safety (2013), 'Road Accidents Statistics in India', *National Crime Records Bureau*, Ministry of Road Transport and Highway, Law Commission of India.

[5] K.Sensarma, H.Singh, N.K. Sharma, N.Balani, S.S. Ravat (2008), 'Leading Causes of Death, 2004 and 2030 (Table-11)', *Road Accident in India -2008*, Ministry of Road Transport and Highways Government of India, New Delhi.

[6] Expert System Team, (2020), 'What is Machine Learning? A definition'

[7] Ian Jack, (2018), 'India has 600 million young people-and they're set to change our world', *The Guardian*

[8] Savolainen, T. P., Mannering, L. F., Lord D., Quddus A.M., (2011), 'The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives', Accident Analysis, Prev. 43, pp.1666-1676.

[9] Mannering, F.L., Shankar V., Bhat C.R, (2016), 'Unobserved heterogeneity and the statistical analysis of highway accident data', Anal. Meth. Accident Res. 11, pp.1-16.

[10] Han, J., Kamber, M., Pei, J., (2012), 'Data mining concepts and techniques. The Morgan Kaufmann Series in Data Management Systems, Third ed. Morgan Kaufmann Publishers. Waltham. MA.

[11] Laird statistics, 'statistics.lared.com', https:// statistics.lared.com/aboutus.php

[12] Ajit Goswami, Ripunjoy Sonwal, (2011), 'A statistical Analysis of Road Traffic Accidents in Dibrugarh City, Assam, India', *Researchgate, InterStat*, ISSN 1941-689X, pp.1-14.

[13] Odero W (1995), 'Road traffic accidents in Kenya: An epidemiological appraisal', East African J, 72 (5), pp. 299 – 305.

[14] Persson A (2008), 'Road traffic accidents in Ethiopia: magnitude, causes and possible Interventions', *Advances in Transportation Studies an international J*, 15: pp. 5-16.

[15] Hossain QS, Adhikary SK, Ibrahim WHW, Rezaur RB (2005), 'Road Traffic Accident The situation in Khulna City, Bangladesh. Proceedings of the Eastern Asia Society for Transportation Studies', 5, pp. 65 – 74.

[16] Jha N, Agrawal CS (2004), 'Epidemiological Study of Road Traffic Accident Cases: A A study from Eastern Nepal', *Regional Health Forum*, 8 (1), pp.15 – 22.

[17] Sarangi L, Parhi L, Parida RK, Panda P (2009), 'A Study on Epidemiological Factors Associated with Road Traffic Accidents Presenting to the Casualty of a Private Hospital in Bhubaneswar', *J of Community Medicine*, 5(2).

[18] Shrinivas PLL,(2004), 'Studies undertaken to identify critical causes of accidents in the highways of Tamil Nadu', *Indian Highways*, 31: pp.11-22.

[19] ROAD ACCIDENT SCENARIO IN INDIA AND ABROAD, Inception report-National Highway Authority of India.

[20] World Health Organization Report on "INJURIES and VIOLENCE", pp 1